A consistent bootstrap linearity test for a regression model with an imprecise response

Ferraro, Maria Brigida
Sapienza University of Rome, Department of Statistical Sciences
P.le A. Moro 5
00185 Rome,Italy
E-mail: mariabrigida.ferraro@uniroma.it

Colubi, Ana
Oviedo University, Department of Statistics
c/ Calvo Sotelo s/n
33007 Oviedo, Spain
E-mail: colubi@uniovi.es

González-Rodríguez, Gil
Oviedo University, Department of Statistics
c/ Calvo Sotelo s/n
33007 Oviedo, Spain
E-mail: gil@uniovi.es

Introduction

In the last years different statistical procedures with fuzzy random variables have been introduced. In particular, in literature there are different works about the regression in this context. A useful kind of fuzzy numbers used for the formalization of imprecise values is the so-called LR family. A linear regression model with an LR fuzzy response and a real explanatory variable has been introduced and analyzed in Ferraro *et al.* (2010, 2011). The main idea is to jointly consider three regression models involving the center and two transformations of the left and the right spread of the fuzzy response variable. In this way it is possible to overcome the non-negativity condition of the spreads that is one of the main difficulties in this context.

Since the inferences developed for such model are meaningful only if the relationship is indeed linear, in this paper a linearity test for the above linear regression model is introduced and discussed.

The proposed linearity test takes inspiration from Stute (1997). It is based on empirical processes of the regressors marked by the residuals. Taking into account some properties of the linear combinations, it can be used a linear combination of the test statistics referred to each model or a test statistic of a model in which the response is a linear combination of the three responses.

Preliminaries

A fuzzy set $\widetilde{A}$ is characterized by means of a membership function $\mu_{\widetilde{A}} : \mathbb{R} \to [0,1]$ so that $\mu_{\widetilde{A}}(x)$ is the membership degree of x in the fuzzy set $\widetilde{A}$ (Zadeh, 1965). The members of the LR family, $\mathcal{F}_{LR}$, are the so-called LR fuzzy numbers, determined by three values: the center, the left and the right spread (see, for example, Coppi *et al.*, 2006). Namely, a mapping $s : \mathcal{F}_{LR} \to \mathbb{R}^3$, i.e., $s(\widetilde{A}) = s_{\widetilde{A}} = (A^m, A^l, A^r)$ (where A^m, $A^l \geq 0$, $A^r \geq 0$ are, respectively, the center, the left and the right spread), is associated to each LR fuzzy set $\widetilde{A}$. In what follows it is indistinctly used $\widetilde{A} \in \mathcal{F}_{LR}$ or

$(A^m, A^l, A^r) \in \mathbb{R}^3$. The membership function of $\widetilde{A} \in \mathcal{F}_{LR}$ can be written as

$$(1) \quad \mu_{\widetilde{A}}(x) = \begin{cases} L\left(\frac{A^m - x}{A^l}\right) & x \leq A^m, \quad A^l > 0, \\ 1_{\{A^m\}}(x) & x \leq A^m, \quad A^l = 0, \\ R\left(\frac{x - A^m}{A^r}\right) & x > A^m, \quad A^r > 0, \\ 0 & x > A^m, \quad A^r = 0, \end{cases}$$

where the functions L and R are particular decreasing shape functions from $\mathbb{R}^+$ to $[0,1]$ such that $L(0) = R(0) = 1$ and $L(x) = R(x) = 0, \forall x \in \mathbb{R} \setminus [0,1]$, and 1_I is the indicator function of a set I. $\widetilde{A}$ is a *triangular* fuzzy number if $L(z) = R(z) = 1 - z$, for $0 \leq z \leq 1$.

The arithmetics considered in $\mathcal{F}_{LR}$ are the natural extensions of the Minkowski sum and the product by a positive scalar for intervals. In details, the sum of $\widetilde{A}$ and $\widetilde{B}$ in $\mathcal{F}_{LR}$ is the LR fuzzy number $\widetilde{A} + \widetilde{B}$ so that $(A^m, A^l, A^r) + (B^m, B^l, B^r) = (A^m + B^m, A^l + B^l, A^r + B^r)$, and the product of $\widetilde{A} \in \mathcal{F}_{LR}$ by a positive scalar γ is $\gamma(A^m, A^l, A^r) = (\gamma A^m, \gamma A^l, \gamma A^r)$. Yang and Ko (1986) have defined a distance between two LR fuzzy numbers $\widetilde{A}$ and $\widetilde{B}$ as follows

$$D^2_{LR}(\widetilde{A}, \widetilde{B}) = (A^m - B^m)^2 + [(A^m - \lambda A^l) - (B^m - \lambda B^l)]^2 + [(A^m + \rho A^r) - (B^m + \rho B^r)]^2,$$

where the parameters $\lambda = \int_0^1 L^{-1}(\omega)\mathrm{d}\omega$ and $\rho = \int_0^1 R^{-1}(\omega)\mathrm{d}\omega$ are related to the shape of the membership function. In the triangular case, $\lambda = \rho = \frac{1}{2}$ (see, for more details, Yang and Ko, 1986). In order to embed the space $\mathcal{F}_{LR}$ into $\mathbb{R}^3$ by preserving the metric a generalization of the Yang and Ko metric has been derived in Ferraro *et al.* (2010). Namely, given $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3) \in \mathbb{R}^3$, it is

$$D^2_{\lambda\rho}(a, b) = (a_1 - b_1)^2 + ((a_1 - \lambda a_2) - (b_1 - \lambda b_2))^2 + ((a_1 + \rho a_3) - (b_1 + \rho b_3))^2,$$

where λ, $\rho \in \mathbb{R}^+$.

According to Puri & Ralescu's sense, the concept of fuzzy random variable (FRV) can be introduced. Let $(\Omega, \mathcal{A}, P)$ be a probability space, a mapping $\widetilde{X} : \Omega \to \mathcal{F}_{LR}$ is an LR FRV if the s-representation of $\widetilde{X}$, $(X^m, X^l, X^r) : \Omega \to \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}^+$ is a random vector (Puri and Ralescu, 1986). The expectation of an LR FRV $\widetilde{X}$ is the LR fuzzy set $E(\widetilde{X}) = (E(X^m), E(X^l), E(X^r))$ and the variance of $\widetilde{X}$ is defined as $\sigma^2_{\widetilde{X}} = var(\widetilde{X}) = E[D^2_{LR}(\widetilde{X}, E(\widetilde{X}))]$ (see, for more details, Ferraro *et al.*, 2010).

Model

Consider a random experiment in which an LR fuzzy response variable $\widetilde{Y}$ and a real explanatory variable X are observed on n statistical units, $\{\widetilde{Y}_i, X_i\}_{i=1,\ldots,n}$. Since $\widetilde{Y}$ is characterized by three real-valued random variables (Y^m, Y^l, Y^r), the regression model proposed in Ferraro *et al.* (2010) concerns this tuple. Due to some difficulties entailed by the non-negativity condition of Y^l and Y^r, the authors proposed modelling the center, a transformation of the left spread and a transformation of the right spread of the response through simple linear regressions (on the explanatory variable X). This can be represented in the following way, letting $g : (0, +\infty) \longrightarrow \mathbb{R}$ and $h : (0, +\infty) \longrightarrow \mathbb{R}$ be invertible:

$$(2) \quad \begin{cases} Y^m = a_m X + b_m + \varepsilon_m, \\ g(Y^l) = a_l X + b_l + \varepsilon_l, \\ h(Y^r) = a_r X + b_r + \varepsilon_r, \end{cases}$$

where ε_m, ε_l and ε_r are real-valued random variables with $E(\varepsilon_m|X) = E(\varepsilon_l|X) = E(\varepsilon_r|X) = 0$. The variance of the explanatory variable X will be denoted by σ^2_X and Σ will stand for the covariance matrix of $(\varepsilon_m, \varepsilon_l, \varepsilon_r)$, whose variances are strictly positive and finite. In the sequel we will assume the existence of all population variances and covariances involved in the developments.

In general, an LR fuzzy random variable $\widetilde{Y}$ and a (real-valued) random variable X can also be related by means of a nonparametric model. As in (2) we jointly consider three equations in which the response variables are the center Y^m and two transformations of the left and the right spreads ($g(Y^l)$ and $h(Y^r)$) of $\widetilde{Y}$, that is,

$$(3) \quad \begin{cases} Y^m = f_m(X) + \varepsilon_m, \\ g(Y^l) = f_l(X) + \varepsilon_l, \\ h(Y^r) = f_r(X) + \varepsilon_r. \end{cases}$$

To estimate model (2), a least squares (LS) approach has been employed. It can be shown that the LS estimators for the parameters of model (2) are strongly consistent and their expressions in terms of the sample moments are as follows

$$\widehat{a}_m = \frac{\widehat{\sigma}_{XY^m}}{\widehat{\sigma}^2_X}, \quad \widehat{a}_l = \frac{\widehat{\sigma}_{Xg(Y^l)}}{\widehat{\sigma}^2_X}, \quad \widehat{a}_r = \frac{\widehat{\sigma}_{Xh(Y^r)}}{\widehat{\sigma}^2_X}, \quad \widehat{b}_m = \overline{Y^m} - \widehat{a}_m \overline{X}, \quad \widehat{b}_l = \overline{g(Y^l)} - \widehat{a}_l \overline{X}, \quad \widehat{b}_r = \overline{h(Y^r)} - \widehat{a}_r \overline{X}.$$

Linear Combinations of variables

In this section some properties of the linear combinations are checked in order to use them in constructing a linearity test.

Proposition 1 *Consider k models f_1, f_2,...,f_k. If there exist k linear combinations of these models that are linear in X:*

$$(4) \quad \sum_{i=1}^{k} w_{ji} f_i = a_j X + b_j, \qquad j = 1, ..., k,$$

with $\sum_{i=1}^{k} w_{ji} f_i$, then there exist $w_j = \sum_{t=1}^{k} \lambda_t w_{tj}$, for $j = 1, ..., k$, such that $\sum_{j=1}^{k} w_j = 1$ and

$$(5) \quad \sum_{j=1}^{k} w_j f_j = aX + b.$$

Remark 1 *If there exist k linear combinations of the models f_j, $\forall j = 1, ..., k$, all the linear combinations of the k models are linear. In particular, if $w_j = 1$ and $w_t = 0$ $\forall t \neq j$ then f_j is a linear model.*

Proposition 2 *If f_j, for $j = 1, ..., k$, is no linear then each linear combination of f_j is almost sure no liner.*

Linearity test

The goal of this section is to test

$$(6) \quad H_0 : \begin{cases} f_m(X) = a_m X + b_m \\ f_l(X) = a_l X + b_l \\ f_r(X) = a_r X + b_r \end{cases}$$

against the alternative

$H_1 : f_m(X), f_l(X), f_r(X)$ are smooth and non-linear functions.

The linearity test proposed in Stute (1997) is based on the integrated mean. The statistic test for a simple regression model, where the response is indicated by Y, is

$$T_n^S = \int_t \left(n^{-1/2} \sum_{i=1}^{n} 1_{X_i \leq t} \left[Y_i - \widehat{Y}_i \right] \right)^2 F(dt),$$

where $\widehat{Y}_i$, for $i = 1, .., n$, are the values of the response predicted by the linear model.

There are two options:

1. To consider a linear combination of each T_n^S calculated on each regression model in (2)
2. To consider the test statistic T_n^S of a regression model in which the response is a linear combination of the three responses in (2), that is,

$$Y = w_m Y_m + w_l g(Y^l) + w_r h(Y^r) = aX + b + \varepsilon.$$

Bootstrap approach

In this context we use a residual-based bootstrap approach. In order to achieve the correctness of the bootstrap test it is necessary to re-sample both the estimated residuals as well as the explanatory variable independently each other (as the conditional distribution of the residuals is considered to be independent of X). This leads to a quite interesting combination for the bootstrap procedure: as the estimated residuals are related with the original ones ε_m, ε_l, ε_r and the explanatory variable X, then the vector $(X, \varepsilon_m, \varepsilon_l, \varepsilon_r)$ is indirectly jointly re-sampled independently and an additional X has to be re-sampled independently of this vector.
In general, we consider the residuals

$$\begin{aligned} \widehat{\varepsilon}_{mi} &= Y_i^m - \widehat{a}_m X_i - \widehat{b}_m, \\ \widehat{\varepsilon}_{li} &= g(Y_i^l) - \widehat{a}_l X_i - \widehat{b}_l, \\ \widehat{\varepsilon}_{ri} &= h(Y_i^r) - \widehat{a}_r X_i - \widehat{b}_r. \end{aligned}$$

We draw a bootstrap sample of the form

$$\left\{ \left(X_i, Y_{1i}^* = \widehat{a}_m X_i^0 + \widehat{b}_m + \widehat{\varepsilon}_{mi}^*, Y_{2i}^* = \widehat{a}_l X_i^0 + \widehat{b}_l + \widehat{\varepsilon}_{li}^*, Y_{3i}^* = \widehat{a}_r X_i^0 + \widehat{b}_r + \widehat{\varepsilon}_{ri}^* \right) \right\}_{i=1,\ldots,n}.$$

In details, $\{(\widehat{\varepsilon}_{mi}^*, \widehat{\varepsilon}_{li}^*, \widehat{\varepsilon}_{ri}^*)\}_{i=1,\ldots,n}$ is an i.i.d. sample from the empirical distribution function of the residuals (see, for more details, Efron & Tibshirani, 1993), that is, it is constructed on the basis of $\{(X_i^*, \varepsilon_{mi}^*, \varepsilon_{li}^*, \varepsilon_{ri}^*)\}_{i=1,\ldots,n}$, sampled from $\widehat{F}_{n,X,\varepsilon_m,\varepsilon_l,\varepsilon_r}$, and $\left\{(X_i^0)\right\}_{i=1,\ldots,n}$ is an i.i.d sample from the empirical distribution function $\widehat{F}_{n,X,c} = \widehat{F}_{n,X} + A$, with $A \to 0$ as $n \to \infty$. The bootstrap statistic is

$$\begin{aligned} T_n^{S*} &= \alpha_m \int_t \left(n^{-1/2} \sum_{i=1}^{n} 1_{X_i^0 \leq t} \left[Y_{1i}^* - \widehat{Y}_{1i}^* \right] \right)^2 F(dt) + \alpha_l \int_t \left(n^{-1/2} \sum_{i=1}^{n} 1_{X_i^0 \leq t} \left[Y_{2i}^* - \widehat{Y}_{2i}^* \right] \right)^2 F(dt) \\ &+ \alpha_r \int_t \left(n^{-1/2} \sum_{i=1}^{n} 1_{X_i^0 \leq t} \left[Y_{3i}^* - \widehat{Y}_{3i}^* \right] \right)^2 F(dt), \end{aligned}$$

where $\widehat{Y}_{1i}^* = \widehat{a}_m^* X_i^0 + \widehat{b}_m^*$, $\widehat{Y}_{2i}^* = \widehat{a}_l^* X_i^0 + \widehat{b}_l^*$ and $\widehat{Y}_{3i}^* = \widehat{a}_r^* X_i^0 + \widehat{b}_r^*$, for $i = 1, ..., n$.

In order to prove the consistency of the residual approach it is necessary to prove that the residual-based bootstrap test statistic has the same asymptotic distribution T_∞^S of

$$T_n^S = \alpha_m \int_t \left(n^{-1/2} \sum_{i=1}^n 1_{X_i \leq t} \left[Y_i^m - \widehat{Y}_i^m \right] \right)^2 F(dt) + \alpha_l \int_t \left(n^{-1/2} \sum_{i=1}^n 1_{X_i \leq t} \left[g(Y_i^l) - \widehat{g(Y_i^l)} \right] \right)^2 F(dt)$$
$$+ \alpha_r \int_t \left(n^{-1/2} \sum_{i=1}^n 1_{X_i \leq t} \left[h(Y_i^r) - \widehat{h(Y_i^r)} \right] \right)^2 F(dt).$$

under the null hypothesis of linearity.

Proposition 3 *Under the assumptions of model (2) and the hypothesis of linearity, if* $E(X^4) < \infty$, $E(\varepsilon_m^4) < \infty$, $E(\varepsilon_l^4) < \infty$, $E(\varepsilon_r^4) < \infty$, *as* $n \to \infty$, *the asymptotic distribution of the bootstrap statistic* T_n^* *is almost surely* T_∞^S.

The application of the bootstrap test based on Proposition 3 is presented in the following algorithm.

Bootstrap Algorithm

Step 1: Compute the values $\widehat{a}_m$, $\widehat{a}_l$, $\widehat{a}_r$, $\widehat{b}_m$, $\widehat{b}_l$ and $\widehat{b}_r$.

Step 2: Compute the residuals $\widehat{\varepsilon}_{mi}$, $\widehat{\varepsilon}_{li}$ and $\widehat{\varepsilon}_{ri}$.

Step 3: Generate a bootstrap sample of the form

$$\left\{ \left(X_i, Y_{1i}^* = \widehat{a}_m X_i^0 + \widehat{b}_m + \widehat{\varepsilon}_{1i}^*, Y_{2i}^* = \widehat{a}_l X_i^0 + \widehat{b}_l + \widehat{\varepsilon}_{2i}^*, Y_{3i}^* = \widehat{a}_r X_i^0 + \widehat{b}_r + \widehat{\varepsilon}_{3i}^* \right) \right\}_{i=1,\ldots,n},$$

and compute the value of the bootstrap statistic T_n^{S*}.

Step 4: Repeat Step 3 a large number B of times to get a set of B estimators, denoted by

$$\{T_{n1}^*, ..., T_{nB}^*\}.$$

Step 5: Approximate the bootstrap p-value as the proportion of values in $\{T_{n1}^*, ..., T_{nB}^*\}$ being greater than T_n.

Simulation studies and concluding remarks

In order to illustrate the empirical significance of the bootstrap test we have used a simulation study. We have considered $B = 1000$ replications of the bootstrap test and 10000 iterations of the test at three different nominal significance levels $\alpha = .01$, $\alpha = .05$, $\alpha = .1$ for different sample sizes n, from 30 to 200. It results that even for small sample sizes the empirical percentages of rejection are quite close to the nominal ones.

We have analyzed the power of the proposed two tests: the linear combination of the Stute's type tests (LCS) and the Stute's type test of the linear combination of responses (SLC). We have considered a population constructed as follows

- $Y_m = 3X + 5 + c_m X^2 + \varepsilon_m$,
- $Y_2 = g(Y_l) = 1.5X + 3.4 + c_l X^2 + \varepsilon_l$,

- $Y_3 = h(Y_r) = 2X + 4.2 + c_r X^2 + \varepsilon_r$,

where X has been drawn as $U(-2, 2)$ and ε_m, ε_l, ε_r as $N(0, 1)$. In this case c_m, c_l and c_r represent the influence of the quadratic component, X^2, on the responses. As the values of the parameters c_m, c_l and c_r get large the models tend to the alternative hypothesis so the percentages of rejection approximate the power of the test. When $c_m = c_l = c_r = 0$ the quadratic component for the three models is null, hence it represents the hypothesis of linearity.

The percentage of times that H_0 is rejected for increasing values of c_m tends to 100, that is, the power tends to 1. This is more evident when we consider two parameters, for example c_m and c_l, getting large.

The test statistic based on a linear combination of T_n^S have higher power than the other test. This is due to the loss of information in considering a linear combination of the responses.

REFERENCES

Bickel, P.J., Freedman, D.A. 1981. Some asymptotic theory for the bootstrap. The Annals of Statistics 9, 1196–1217.

Coppi R, D'Urso P, Giordani P, Santoro A (2006) Least squares estimation of a linear regression model with LR fuzzy response. Comput Statist Data Anal 51:267–286

Efron, B., Tibshirani, R.J. 1993. An introduction to the bootstrap, Chapman & Hall, New York.

Ferraro, M.B., Coppi, R., González-Rodríguez, G., Colubi, A., 2010. A linear regression model for imprecise response. International Journal of Approximate Reasoning 51, 759–770.

Ferraro, M.B., Colubi, A., González-Rodríguez, G., Coppi, R., 2011. A determination coefficient for a linear regression model with imprecise response. Environmetrics (in press, doi: 10.1002/env.1056).

González-Rodríguez, G., Blanco, A., Colubi, A., Lubiano, M.A., 2009. Estimation of a simple linear regression model for fuzzy random variables. Fuzzy Sets and Systems 160, 357–370.

González-Rodríguez, G., Colubi, A., Gil, M.A., 2010. Fuzzy data treated as functional data: A one-way ANOVA test approach. Computational Statistics and Data Analysis (in press, doi:10.1016/j.csda.2010.06.013).

Puri, M.L., Ralescu, D.A., 1986. Fuzzy random variables. Journal of Mathematical Analysis and Applications 114, 409–422.

Stute, W., 1987. Nonparametric model checks for regression. Ann Statist 25:613–641

Stute, W., Gonzalez Manteiga, W., Presedo Quindimil, M., 1998. Bootstrap approximations in model checks for regression. J Amer Stat Assoc 93:141–149

Yang, M.S., Ko, C.H., 1996. On a class of fuzzy c-numbers clustering procedures for fuzzy data. Fuzzy Sets Syst 84:49–60

Zadeh, L.A., 1965. Fuzzy sets. Information and Control 8, 338–353.